DOCUMENT RESUME

ED 477 357 TM 034 919

AUTHOR	Rudner, Lawrence M., Ed.; Schafer, William D., Ed.
TITLE	Practical Assessment, Research and Evaluation, 2001.
INSTITUTION	ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.; Maryland Univ., College Park. Dept. of Measurement, Statistics & Evaluation.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
ISSN	ISSN-1531-7714
PUB DATE	2001-12-00
NOTE	33p.; "Practical Assessment, Research & Evaluation" is an electronic-only journal covered on an article-by-article basis in "Current Index to Journals in Education" (CIJE). For the first 22 articles in volume 7, see ED 458 254. For articles 23 through 26, see TM 525 154-157.
AVAILABLE FROM	For full text: http://ericae.net/pare.
PUB TYPE	Collected Works - Serials (022) -- ERIC Publications (071)
JOURNAL CIT	Practical Assessment, Research and Evaluation; v7 n23-26 2001
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Educational Research; Elementary Secondary Education; *Essays; *High Stakes Tests; *Intelligence Tests; *Scoring Rubrics; *Timed Tests

ABSTRACT

This document consists of articles 23 through 26 published in the electronic journal "Practical Assessment, Research & Evaluation" in 2001: (23) "Effects of Removing the Time Limit on First and Second Language Intelligence Test Performance" (Jennifer Mullane and Stuart J. McKelvie); (24) "Consequences of (Mis)use of the Texas Assessment of Academic Skills (TAAS) for High-Stakes Decisions: A Commentary on Haney and the Texas Miracle in Education" (J. Thomas Kellow and Victor L. Willson); (25) "Designing Scoring Rubrics for Your Classroom" (Craig A. Mertler); and (26) "An Overview of Three Approaches to Scoring Written Essays by Computer" (Lawrence Rudner and Phil Gagne). (SLD)

Volume 7, 2001

ISSN 1531-7714

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal

(Articles 23-26)

Practical Assessment, Research and Evaluation
is listed among the ejournals in education at the website for the
AERA SIG "Communications Among Researchers"

Practical Assessment, Research and Evaluation (PARE) is an on-line journal published by the **ERIC Clearinghouse on Assessment and Evaluation (ERIC/AE) and the Department of Measurement, Statistics, and Evaluation at the University of Maryland, College Park.** Its purpose is to provide education professionals access to refereed articles that can have a positive impact on assessment, research, evaluation, and teaching practice, especially at the local education agency (LEA) level.

Manuscripts published in *Practical Assessment, Research and Evaluation* are scholarly syntheses of research and ideas about issues and practices in education. They are designed to help members of the community keep up-to-date with effective methods, trends and research developments. While they are most often prepared for practitioners, such as teachers, administrators, and assessment personnel who work in schools and school systems, *PARE* articles can target other audiences, including researchers, policy makers, parents, and students.

Manuscripts to be considered for *Practical Assessment, Research and Evaluation* should be short, 2000-8000 words or about eight pages in length, exclusive of tables and references. They should conform to the stylistic conventions of the American Psychological Association (APA). See the Policies section of this web site for technical specifications and a list of suggested topics. Manuscripts should be submitted electronically to pare2@ericae.net. Articles appearing in *Practical Assessment, Research and Evaluation* also become available in the ERIC database through the ERIC Digest Series. Many articles published in *PARE* were previously published as part of the *ERIC/AE Digest Series.*

Practical Assessment, Research and Evaluation is listed among the ejournals in education by the Scholarly Publishing and Academic Resources Coalition(SPARC), and the website for the AERA SIG "Communications Among Researchers".

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title □ Go

Mullane, Jennifer & Stuart J. McKelvie (2001). Effects of removing the time limit on first and second language intelligence test performance. *Practical Assessment, Research & Evaluation*, 7(23). Available online: http://ericae.net/pare/getvn.asp?v=7&n=23. This paper has been viewed 4761 times since 11/6/01.

▸ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

▸ Find articles in ERIC written by
Mullane, Jennifer
Stuart J. McKelvie

Effects of Removing the Time Limit on First and Second Language Intelligence Test Performance

Jennifer Mullane and Stuart J. McKelvie
Department of Psychology
Bishop's University

Abstract

Canadian post-secondary students with a moderate level of second language competence in English or French took the Wonderlic Personnel Test with the standard 12-min time limit or with no time limit. Participants who took the timed test in their second language scored lower than those who took it in their mother tongue, but the disadvantage was greater for limited French-proficient (LFP) students than for limited English-proficient (LEP) students. Scores increased with no time limit and the gain was greater on the French test for LFP students than for mother tongue students. On the English test, the gain was similar for LEP students and mother tongue students. It is concluded that the time accommodation can be applied to clients who are taking an intelligence test in their second language.

Maximum performance psychological tests enjoy widespread use in North America, particularly for educational and employment decisions. Ever since they were administered en masse to U.S. army recruits and immigrants in the early 1900s, the testing of minorities, particularly limited-English proficient (LEP) clients, has been controversial (Samelson, 1977). The issue is test fairness or "intercept bias": the possibility that the minority group scores lower than the majority on the test but not on the criterion (Anastasi & Urbina, 1997). LEP clients may be disadvantaged on a standardized intelligence test given in English, but they may be just as capable in school or on the job as native English speakers (Cummins, 1987).

In the U.S., 15 to 20% of students speak a language other than English at home (Geisenger & Carlson, 1992). In Canada, the corresponding percentage is 32 (23% speaking the second official language, French) (Statistics Canada, 1996). Although it may (Angus, 2000) or may not (Lam, 1993; Rivera, Vincent, Hafner & LaCelle-Peterson, 1997) be appropriate to administer achievement tests to LEP clients in the majority language, doing so with tests of general cognitive functioning (intelligence) is questionable. Test scores of LEP clients will vary with their English skills, and are likely to underestimate their ability to learn, which should remain relatively constant (Angus, 2000). Indeed, even bilingual children perform more poorly than monolinguals on standardized tests (Valdes & Figueroa, 1994).

To alleviate this problem, students could be assessed with a nonverbal test of intelligence, with a test in their own language, or with a modified test in the majority language. Although nonverbal tests have been defended (e.g., Bracken & M^{c}Callum, 2001), they may not measure the same cognitive functions as verbal intelligence tests (Angus, 2000; Anastasi & Urbina, 1997), and they are less successful than verbal tests at predicting educational performance (Angus, 2000; Geisenger & Carlson, 1992; Gregory, 2000). Such problems can be avoided by administering the test in the client's own language, but there can be problems in translation, with no guarantee that the second-language version measures the same construct as the original (Wonderlic, 1992). Moreover, few tests have been translated from English (Geisenger & Carlson, 1992).

The third solution is to modify the original test or test conditions to minimize the disadvantage to the LEP client. For example, grammar may be simplified, a glossary of terms may be provided, or time constraints may be relaxed (Azar, 1999; Rivera et al., 1997). In a survey of statewide assessment programs, it was found that 52% of states permitted amendments for LEP students, and the most popular one (81%) was giving extra time. As with test translation, there is no guarantee that such changes will preserve test validity, but an accommodation can be said to work if it provides a "differential advantage" to the LEP client (Azar, 1999). This means that the gain that accrues from the accommodation must be greater for LEP than for English mother tongue clients. It has been found that SAT scores increase for LEP clients who have extra time, but it is not clear if the gain is greater than for other regular students (Azar, 1999). More research is needed on the effect of test modifications (Rivera et al., 1997).

The purpose of the present study is to investigate the effect of removing the time limit for students taking a standardized intelligence test in their second language. To evaluate whether the accommodation worked, the results were compared with those obtained by students who took the test in their mother tongue. In our study, participants completed the 50-item spiral-omnibus Wonderlic Personnel Test (WPT) with the standard time limit or with unlimited time. The WPT is widely employed in personnel selection, and can be used in educational settings (Wonderlic, 1992). Although Murphy (1984) doubts the manual's claim that it measures the "ability to learn," he states that it is a useful predictor of job performance. Furthermore, it can be classified as a test of general intelligence, because the items are based on the Otis Self-administering Tests of Mental Ability, and cover numerical reasoning, verbal reasoning, synonyms-antonyms, nonverbal reasoning, information, and attention to detail (McKelvie, 1992). Moreover, its loadings on Aptitude G of the General Aptitude Test Battery exceed .56, and its scores correlate moderately to very highly (up to .90; Dodrill, 1981; Dodrill & Warner, 1988) with the Wechsler Adult Intelligence Scale in general and psychiatric groups. Wonderlic (1992) also states that the test correlates reasonably (from .30 to .45) with academic achievement, so that it can be used as a selection and counseling tool in postsecondary education. Correlations at the university level are lower than those at high school (McKelvie, 1989, 1992), but this may be due to restriction of range.

Under standard instructions, the WPT is administered with a 12-min time limit, which makes it a highly-speeded test (Davou & McKelvie, 1984). Belcher (1992) states that this may be unfair for the increasing number of LEP clients. When answering an item, they may have to translate it into their mother tongue to fully understand its meaning. By spending more time than the native English speaker on item comprehension, they will not be able to complete as many items in the restricted time allowed. Although they may answer correctly on the items that they attempt, their obtained score will underestimate their true score. Indeed, the Wonderlic (1992) manual reports that

Hispanic Americans scored about six points lower than white Americans on the English WPT. It is also notable that one study has shown that the time accommodation worked on the WPT for students with weak study habits (Davou & McKelvie, 1984). In the standard timed condition, scores were lower for students with weak than with strong study habits. However, the gain with unlimited time was greater for the weak than the strong studiers.

The WPT has also been translated into French for Canadian use, particularly in Québec, where the majority mother tongue is French. Here, English speakers are the minority. On the other hand, in the Québec academic setting where testing took place, the language of instruction is English and students are enrolled from all parts of Canada (as well as the U.S. and overseas). Here, French speakers are the minority. Thus, we had the opportunity to examine the effect of removing the time limit not only for LEP (French mother tongue) clients taking the test in English, but also for LFP (limited-French proficient, English mother tongue) clients taking the test in French. It was hypothesized that, although all participants would benefit from the relaxed time constraint, the gain would be greater for people taking the test in their second language. That is, the gain should be greater on the English WPT for French (LEP) compared to English mother tongue students, and on the French WPT for English (LFP) compared to French mother tongue students.

Method

Participants

The participants were 133 (89 women, 44 men; mean age 21.6 yr.) postsecondary college and university students who reported English or French as their mother tongue and at least moderate competence in the other language. After matching for gender, English and French first-language participants were assigned randomly to one of four experimental conditions resulting from two levels of two independent variables: timing (Limited Time, Unlimited Time) and test language (English, French).

Materials and Procedure

As described above, the Wonderlic Personnel Test (WPT) measures general intelligence. Its reliability has been estimated as .84 to .94 (test-retest, even for a 5-yr, period; Dodrill, 1983), .88 to .94 (split-half; Wonderlic, 1992), and .73 to .95 (alternate-form; Schoenfeldt, 1985).

Although participants reported that they had moderate competence in their second language, this was assessed more formally with two written tests of second-language proficiency: a 13-item self-report measure (Self Evaluation Questionnaire) in their mother tongue, and a 17-item objective test (Objective Cloze Test) in their second language. The Objective Cloze Test (OCT) was developed from an experimental version of the Test of English as a Foreign Language (TOEFL) (Hale, Stansfield, Rock, Hicks, Butler, & Oller, 1989). In the cloze procedure, words in a text are deleted and must be supplied by the examinee. In the OCT, they were given four choices from which to choose. Scores from multiple-choice cloze tests correlate highly with those from fill-in-the-blank versions (Chappelle & Abraham, 1990).

According to Al-Fallay (1997), the cloze procedure has concurrent validity as a predictor of other second-language achievement tests, but its face validity is questionable. We followed their recommendation to supplement the cloze test with another technique, and chosc sclf-asscssment, which has been useful (Ross, 1998). The Self Evaluation Questionnaire (SEQ) was adapted from the Bishop's University ESL (English as a Second Language) department's screening questionnaire. Items tapped various aspects of language ability and were answered on a 7-point Likert scale. Examples are: "When I speak French among a small group of people that I know well, I feel 1 (uneasy) to 9 (very much at ease)"; "I can understand newspaper articles without the use of a dictionary: 1 (not at all) to 7 (perfectly)". Both the OCT and SEQ were constructed in English and translated professionally into French.

Participants were tested individually or in small groups. After signing a consent form, they completed the SEQ then the OCT. The SEQ was given first to avoid contamination of self-reports by perceived performance on the objective test. Following these tests, the WPT was administered then participants were debriefed. The WPT was scored using the standard key for the English version. However, for the French version, two adjustments had to be made because of imprecise translation.

Results and Discussion

Second-Language Competence

The correlation between OCT and SEQ scores for all 133 participants was .529, p < .01. This indicates that the two tests were related, but tapped slightly different aspects of second-language competence, supporting the view that it should be assessed with more than one technique (Al-Fallay,1997).

In the present design, second-language competence should be moderate and similar in the different experimental conditions. If participants were perfectly bilingual (scoring close to maximum), second-language testing would not an issue, and if they were essentially monolingual (scoring close to 0), they would not be taking a test in another second language. To evaluate this, 2 X 2 X 2 (Timing X Mother Tongue X Test Language) factorial ANOVAs were conducted for each second-language test. There was a significant effect of mother tongue for both OCT scores, F(1, 125) = 4.49, p < .05, and for SEQ scores, F(1, 125) = 11.57, p < .01. Mother-tongue French speakers scored higher on the English OCT than mother-tongue English speakers scored on the French OCT (Ms = 12.3. 11.3). French speakers also rated themselves as more proficient in English than English speakers rated themselves in French (Ms = 62.5, 53.6). Although the tests were professionally translated from English to French, the OCT may not be equally difficult in each language. However, the fact that the results agreed with those of the SEQ indicates that our French speakers were more proficient in English than were our English speakers in French. At the same time, none of the mean scores was extreme. On the OCT, the maximum possible score was 17 (Ms were 12.3, 11.3) and on the SEQ it was 91 (Ms were 62.5, 53.6). Moreover, the lack of any significant interactions indicates that second-language competence was matched in the four timing/test language conditions.

Because the major comparisons of interest were the effects of time for people taking the test in their first or second language, the different levels of second-language competence were dealt with by including OCT and SEQ scores as covariates in the analysis of WPT scores.

Intelligence Test Performance

Initially, a 2 X 2 X 2 (Timing X Mother Tongue X Test Language) factorial ANOVA (with SEQ and OCT as covariates) was conducted on WPT scores. Table 1 shows the means in each of the eight conditions. Not surprisingly, there was a significant effect of timing, F(1, 123) = 150.46, p < .001, with higher scores in the unlimited time than in the limited time condition (Ms = 32.58, 22.17). Converted to Cohen's (1977) standardized effect size (d), this difference was 2.10 which clearly exceeds his guideline of 0.80 for a large effect.

Table 1: Mean Wonderlic Personnel Test Scores in Each Condition

	Limited Time			Unlimited Time		
Test Language	n	M	SD	n	M	SD
English						
English Mother Tongue	16	26.33	4.24	17	34.24	6.08
French Mother Tongue (LEP)	18	23.03	5.86	13	30.70	7.50

French						
English Mother Tongue (LFP)	17	16.46	4.62	16	31.94	5.10
French Mother Tongue	18	22.74	4.36	18	33.49	3.07

Note. Maximum score = 50. Means are adjusted for covariates. LEP = limited-English proficient, LFP = limited-French proficient.

There was also a significant effect of test language, F(1, 123) = 9.32, p < .01, with lower scores on the French WPT than on the English WPT (Ms = 26.09, 28.66). However, timing interacted significantly with test language, F(1, 123) = 12.15, p < .01. The gain from the limited to the unlimited time condition was greater on the French test (Ms = 19.42, 32.76; gain = 13.34) than on the English test (Ms = 24.92, 32.39; gain = 7.47). There was also a significant interaction between mother tongue and test language, F(1, 123) = 13.92, p < .01. For people whose mother tongue was English, scores were lower on the French than on the English test (Ms = 24.29, 30.03), but for those whose mother tongue was French, the scores were very similar (Ms = 27.88, 27.28).

These results indicate that the effects of timing and of mother tongue were greater on the French than on the English test. Because it was predicted that the effect of timing would be greater for second- than for first-language participants on each test, planned 2 X 2 (Timing X Mother Tongue) ANOVAs (again with OCT and SEQ as covariates) were conducted separately for the English and French WPTs. For the French version of the test, all three effects were significant: timing, F(1, 63) = 157.18, p < .01, mother tongue, F(1, 63) = 13.15, p < .01, and their interaction, F(1, 63) = 5.04, p < .05. Scores were higher in the unlimited than in the limited time condition, and for French mother tongue than for English mother tongue (LFP) participants. However, and of particular interest here, the latter difference was only significant with limited time. Here, LFP students (English native speakers) performed more poorly than French native speakers, t(63) = 4.24, p < .01, but when the time constraint was removed, they did not, t(63) = 1.07, p > .10. Another way of looking at this is that both groups of participants benefited from extra time, but the gain was greater for the LFP students (15.5 points, d = 3.54) than for the French native speakers (10.8 points, d = 2.44). In other words, the time accommodation worked because it provided a differential advantage to the LFP clients (Azar, 1999).

For the English version of the test, there were significant effect of timing, F(1, 58) = 34.96, p < .01, and of mother tongue, F(1, 58) = 6.31, p < .05, but not their interaction, F(1, 58) = 0.01, p > .90. Scores were higher in the unlimited than in the limited time condition, and for English mother tongue than for French mother tongue (LEP) participants. Thus, although LEP students performed more poorly than English native speakers with the standard 12-min time limit, and although they benefited from unlimited time, the gain was not greater than that for English speakers. In fact, the effect size for time was d = 1.46 (7.7 points) for French speakers and d = 1.50 for English speakers (7.9 points). Here, the time accommodation did not work.

Why did the time accommodation work for LFP but not for LEP participants? The answer may be that, in the timed condition, the LFP disadvantage on the French test (Ms = 16.46, 22.74; difference = 6.28) was greater than the LEP disadvantage on the English test (Ms = 23.03, 26.33; difference = 3.30). The accommodation may only be effective if the disadvantage is great.

But why was the LFP disadvantage greater than the LEP disadvantage? The most obvious reason is that French second-language competence was less than English second-language competence as assessed on the OCT and SEQ. In fact, the mean levels of competence were not extreme, and the differences were controlled via covariance analysis[1]. Perhaps the answer is that the tests did not fully capture the fact that the participants whose second language was English were studying in an English-speaking institution, and who probably had more practice listening, reading and writing in their second language than did participants who second language was French. In fact, the SEQ only had one question about frequency of second-language use, and it only referred to speaking. To aid in the identification of clients likely to perform poorly on the timed WPT in their second language, the SEQ should be expanded to include information about reading and writing. It might also be noted that if the present study had also been conducted in a French-speaking institution, the results might have been reversed. That is, the gain from unlimited time might have been greater for LFP than for LEP students.

Although these results show that the time accommodation can work, they do not show whether the WPT scores obtained with unlimited time are as valid as those obtained with limited time. Wonderlic (1992) himself discusses this issue, stating that "while untimed scores are valid assessments of cognitive ability, they are not as accurate as the timed scores." Notably, a 25-item short form of the WPT given with unlimited time was as reliable as the full version (when corrected for length), and also had a similar criterion validity coefficient for predicting university grades (McKelvie, 1994). However, because his studies indicated that people taking the test under both conditions scored about six points higher with unlimited time, Wonderlic recommends that this score be used to estimate the timed score by subtracting six points from it.

French mother-tongue speakers (LEP participants) scored only slightly lower than English mother-tongue speakers on the English WPT with unlimited time. Because English mother-tongue speakers (LFP) did not score significantly lower than French mother-tongue speakers on the French WPT in this condition, extra time minimized or removed the second language disadvantage. Therefore, we suggest that people taking an intelligence test in their second language be permitted the accommodation of unlimited time. In the case of the WPT, their timed score can then be estimated by subtracting six points.

Conclusion

These results provide experimental evidence that the time accommodation can work for people whose second-language intelligence test limited time score is clearly lower than that of mother-tongue participants, and it does no injustice to those whose limited time score is only slightly lower. Therefore, we recommend removing time limits on standardized intelligence tests for clients taking them in their second language. The present measures of second-language competence should be expanded, and future research should obtain more information about the psychometric properties (norms, reliability, validity) of untimed intelligence tests.

References

Al-Fallay, I. (1997). Investigating the reliability and validity of the fixed ratio multiple-choice cloze test. *Human and Social Sciences, 24*, 507-526.

Anastasi, A., & Urbina, S. (1997). *Psychological testing*, 7th ed. Upper Saddle River, NJ: Prentice-Hall.

Angus, W. A. (2000). Using achievement tests, diagnostic (achievement) tests, and tests of intelligence with ESP populations. http://www.psychtest.com/ESLtest.htm.

Azar, B. (1999). Fairness a challenge when developing special needs tests. *APA Monitor Online, 30.* http://www.apa.org/monitor/dec99/in2.html.

Belcher, M. J. (1992). Review of the Wonderlic Personnel Test. In J. J. Kramer & J. C. Conoley (eds.), *The Eleventh Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.

Bracken, B. & McCallum, R. S. (2001). Assessing intelligence in a population that speaks more than tow hundred languages: A nonverbal solution. In L. A. Suzuki and J. G. Ponterotto (Eds.), *Handbook of multicultural assessment:*

Clinical, psychological, and educational applications, 2nd ed. San Francisco, CA: Jossey-Bass, Inc.

Chapelle, C. A., & Abraham, R. G. (1990)/ Cloze method: What difference does it make? *Language Testing, 7,* 121-146.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Cummins, J. (1997). Psychoeducational assessment in multicultural school systems. *Canadian Journal for Exceptional Children, 3,* 115-117.

Davou, D., & McKelvie, S. J. (1984). Relationship between study habits and performance on an intelligence test with limited and unlimited time. *Psychological Reports, 54,* 367, 371.

Dodrill, C. B. (1981). An economical method for the evaluation of general intelligence in adults. *Journal of Consulting and Clinical Psychology, 49,* 668-673.

Dodrill, C. B. (1983). Long-term reliability of the Wonderlic Personnel Test. *Journal of Consulting and Clinical Psychology, 51,* 316-327.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56,* 145-147.

Edinger, J. D., Shipley, R. H., Watkins, C. E., & Hammett, E. B. (1985). Validity of the Wonderlic Personnel Test as a brief IQ measure in psychiatric patients. *Journal of Consulting and Clinical Psychology, 53,* 937-939.

Geisenger, K. F., & Carlson, J. F. (1992). Assessing language-minority students. Practical Assessment, Research & Evaluation, 3(2). Available online: http://ericae.net/pare/getvn.asp?v=3&n=2.

Gregory, R. J. (2000). *Psychological testing: History, principles, and applications,* 3rd ed. Boston: Allyn and Bacon.

Hale, G., Stansfield, C., Rock, D., Hicks, M., Butler, F., & Oller, J. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6,* 49-78.

Lam, T. C. M. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26,* 179-191.

McKelvie, S. J. (1989). The Wonderlic Personnel Test: Reliability and validity in an academic setting. *Psychological Reports, 65,* 161-162.

McKelvie, S. J. (1992). Does memory contaminate test-retest reliability? *The Journal of Psychology, 119,* 59-72.

McKelvie, S. J. (1994). Validity and reliability findings for an experimental short form of the Wonderlic Personnel Test in an academic setting. *Psychological Reports, 75,* 907-910.

Murphy, K. R. (1984). The Wonderlic Personnel Test. In D. J. Keyser and R. C. Sweetland (Eds.), *Test critiques* (volumes I-VI). Kansas City, MO: Test Corporation of America, pp. 769-775.

Rivera, C., Vincent, C., Hafner, A., & LaCelle-Peterson, M. (1997). Statewide assessment programs: Policies and practicies for the inclusion of limited English proficient students. *Practical Assessment, Research & Evaluation, 5*

(13). Available online: http://ericae.net/pare/getvn.asp?v=3&n=2.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiental factors. *Language Testing, 15,* 1-20.

Samelson, F. (1977). World War I intelligence testing and the development of psychology. *Journal of the History of the Behavioral Sciences, 13,* 274-282.

Schoenfeldt, L. F. (1985). Review of the Wonderlic Personnel Test, In J. V. Mitchell Jr. (ed.). *The ninth mental measurements yearbook.* Vol 2. Lincoln, NE: University of Nebraska Press.

Statistics Canada (1996). 1996 Census figures. http://www.statcan.ca/english/Pgdb/People/Population/demo29d.htm).

Valdes, G., & Figueroa, R. A. (1994). *Bilingualism and testing: A special case of bias.* Stamford, CT: Ablex Publishing Corporation.

Wonderlic, E. F. (1992). *Wonderlic Personnel Test User's Manual.* Libertyville, IL: E. F. Wonderlic.

Footnote

[1] A second analysis of WPT scores was conducted in which the sample size was reduced to 109 by removing participants with higher English and lower French second-language OCT and SEQ scores. The results were the same, with the exception that LEP participants did not score significantly lower than English mother tongue participants on the English WPT. The key point is that, once again, the gain in scores from extra time was only of significant benefit to LFP participants on the French WPT.

Author Note

We thank Dr. Richard Kruk for advice on second language proficiency testing, Denise Bernier for pointing out translation ambiguities, and four reviewers for helpful critical comments. Send correspondence to Stuart J. McKelvie, Department of Psychology, Bishop's University, Lennoxville, Québec J1M 1Z7, Canada. Electronic address is smckelvi@ubishops.ca.

Descriptors: Bilingual Education; Test Format; Evaluation Methods; Intelligence Tests; Language Proficiency; Time Factors [Learning]

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment,
Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title □ Go

Kellow, J. Thomas & Victor L. Willson (2001). Consequences of (mis)use of the texas assessment of academic skills (taas) for high-stakes decisions: a comment on haney and the texas miracle in education. *Practical Assessment, Research & Evaluation*, 7(24). Available online: http://ericae.net/pare/getvn.asp?v=7&n=24. This paper has been viewed 3804 times since 12/7/01.

- Find similar papers in
  ERICAE Full Text Library
  Pract Assess, Res & Eval
  ERIC RIE & CIJE 1990-
  ERIC On-Demand Docs
- Find articles in ERIC written by
  Kellow, J. Thomas
  Victor L. Willson

Consequences of (mis)use of the Texas Assessment of Academic Skills (TAAS) for high-stakes decisions: A comment on Haney and the Texas miracle in education.

J. Thomas Kellow, University of Houston
Victor L. Willson, Texas A&M University

Abstract

This brief paper explores the consequences of failing to incorporate measurement error in the development of cut-scores in criterion-referenced measures. The authors use the case of Texas and the Texas Assessment of Academic Skills (TAAS) to illustrate the impact of measurement error on "false negative" decisions. These results serve as further evidence to support Haney's (2000) contentions regarding the (mis)use of high-stakes testing in the state of Texas.

Walt Haney's (2000) treatise on *The Myth of the Texas Miracle in Education* highlights a number of concerns related to high-stakes decision-making in K-12 settings. His elaboration of the history and development of the Texas Assessment of Academic Skills (TAAS) was illuminating, especially for those unfamiliar with the often capricious fashion in which standards are ultimately set for high-stakes decisions. The evidence Haney presents in evaluating the TAAS fits well with Samuel Messick's (1994) argument for considering the *consequences* of test use. Although sometimes referred to as *consequential validity*, Messick considered this aspect of test interpretation and use to be one of many forms of *construct validity*. His view of the evaluative role of validity is summarized nicely in the following paragraph:

> When assessing any of these constructs – whether attributes of persons or groups, of objects or situations – validity needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability, and fairness. This is so because validity, comparability, and fairness are not just measurement principles, they are *social values* that have meaning and force outside of measurement whenever evaluative judgment and decisions are made. As such, validity assumes both a scientific and political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion or by expert judgments that test content is relevant to the proposed test use (p. 1).

One technical aspect of the TAAS that Haney takes to task is the reliability of scores yielded by the test and the potential for misclassification as a function of measurement error. As the author notes, "The reason the setting of passing scores on a high-stakes test such as the TAAS is so important is that the passing score divides a continuum of scores into just two categories, pass and fail. Doing so is hazardous because all standardized test scores contain some degree of measurement error" (Haney 2000, p. 10). Haney also points out that measures of test-retest (or alternate-form) reliability and not internal consistency should ideally be used to inform judgment as to the potential for misclassfication due to measurement error. In light of the conspicuous absence of test-retest data on the TAAS, he attempts to use extant data to approximate this reliability estimate as compared to the internal consistency (KR20) estimates provided by the Texas Education Agency (TEA). Although the approach he uses is somewhat problematic (see Wainer, 1999), it is clear that these test-retest estimates are lower than the KR20 estimates.

The thrust of Haney's argument is that measurement error inherent in the TAAS (or any other measure, for that matter) contributes appreciably to the rate of "false negatives," or misclassifying *passing* students as *non-passing*. His focus, however, is exclusively on the tenth-grade Exit-Level TAAS since this is the test high school students in the state of Texas must pass in order to graduate from high school (although it is not the sole criterion). But in some districts, the TAAS is also being used at other grade levels for high-stakes decisions – namely, promotion and retention of students. Waco Independent School District and Houston Independent School District, among others, require students in grades 3 through 8 to pass all portions of the TAAS in order to be considered for promotion to grade level. Students who fail any subtest during the statewide spring administration can expect to attend mandatory summer school, after which they are tested on a "released" version of the TAAS. Those students who fail are held back in grade. This stands in contrast to the Exit-Level testing schedule where, as Haney notes, students have as many as eight opportunities to pass the exam. In addition, the TEA has announced plans to use the TAAS for promotion/retention decisions in the third, fifth, and eighth grades beginning in the fall of 2001 (Texas Education Agency, 1999).

Given the current political climate and the clarion calls for school district accountability throughout the nation, the number of districts in Texas using the TAAS for promotion and retention decisions in all tested grades no doubt will increase. Indeed, the Houston model has been lauded for its strict criteria for student grade promotion, which is believed to increase student motivation and achievement while *reducing* the student dropout rate (Markley, 1999). Our purpose in this response is to elaborate on the potential social consequences of using the TAAS for high-stakes decisions, specifically grade promotion and retention. The question we asked is rather simple: "Given the imperfect reliability of the TAAS test, how many students with a true passing score are potentially misclassified as failing across grades and subtests in a given year"? Fortunately, sufficient summary statistics and frequency distributions were available from the TEA to estimate these numbers.

Method

The most recent reliability estimates for the TAAS were reported for the 1998-99 school year; therefore we used means, standard deviations, and frequency distributions for this same year. We want to emphasize that there are a number of ways to go about this estimation process. The method presented emerged because it is fairly intuitive and required minimal summary statistics from the data.

First, we should note that the 70% passing standard on the TAAS mentioned by Haney is *not* fixed across grades and subtests. Because of differential item and thus test difficulty, the proportion of items correct needed to pass a given subtest ranges from .64 to .75, according to the TEA. The Texas Learning Index (TLI) was developed by the TEA for the purpose of, among other things, providing a consistent passing standard across test forms. The TLI is a linear standardized scoring transformation with a standard deviation of 15 and an anchor (rather than mean) of 70, which represents the passing standard for a given subtest at a given grade. The TLI is calculated in z-score form as:

$$\text{TLI}= [(z_{\text{observed}} - z_{\text{passing}})*15] + 70 \qquad (1)$$

Although the TLI observed score of 70 is the passing standard, this standard fails to incorporate measurement error in determining the appropriate cut score. Specifically, the process of modifying cut-scores involves determining the domain score in proportion-correct form that constitutes mastery (τ_0) and then adjusting this value to estimate a new cut score (X_0) in number-correct form that reflects the measurement error in the test data (Crocker & Algina, 1986). Huynh and Sanders (1980) provide an approximate procedure for this purpose that works well when a test consists of 20 or more items and the observed proportion correct cut score falls within .50 to .80. The TAAS subtests meet both criteria. This formula is given as:

$$X_0 = \frac{n - KR_{21}}{KR_{21}} \tau_0 + \frac{KR_{21} - 1}{KR_{21}} \mu_x + .5 \qquad (2)$$

Where *n* is the number of items on the test, τ_0 is the observed proportion-correct cut score, and μ_x is the mean number of items correct. As noted by Crocker and Algina (1986), as KR21 approaches 1.0, X_0 approaches $n\,\tau_0 + .5$ irrespective of the value of μ_x.

Our question focuses on the estimate X_0 when transformed to a TLI score metric. Put simply: What is the passing TLI adjusted for measurement error for a given subtest in a given grade, and what percent and number of students met this adjusted criterion but not the standard cut score of 70? The following steps were employed to determine X_0 in TLI form:

1. calculate X_0 in raw score form;
2. transform X_0 into a z-score;
3. substitute z X_0 for z observed in Formula 1.

Results

Table 1 provides the adjusted cut score X_0 in the TLI metric across grades for both the mathematics and reading subtests.

Table 1
TLI passing cut-scores adjusted for measurement error by subtest and grade

Grade	*Reading*	*Mathematics*
3rd	67.5	67.4
4th	67.2	67.6
5th	67.2	67.0
6th	67.6	68.2
7th	67.9	68.4
8th	67.7	68.3
10th	66.9	68.7

We then used TLI frequency distributions for the 1998-99 administration of the TAAS to determine the percent and number of students who received a TLI score of X_0 or higher but less that 70. Because the TLI frequency distribution tables obtained from TEA report whole number values, we rounded the obtained X_0 estimates to the closest whole number. These data are presented in Table 2 disaggregated by subtest and grade.

Table 2
Percent and number of potentially misclassified students by subtest and grade

Grade	*Reading*	*Mathematics*
3rd	1.7 (n=4,243)	2.9 (n=7,151)
4th	1.7 (n=4,105)	2.1 (n=5,239)
5th	2.2 (n=5,509)	2.4 (n=6,007)
6th	2.2 (n=5,725)	2.5 (n=6,653)
7th	2.0 (n=5,410)	2.8 (n=7,474)
8th	1.4 (n=3,620)	2.6 (n=6,903)
10th	2.9 (n=6,570)	1.6 (n=3,650)

Because of the rounding procedure mentioned earlier, these percentages and student numbers are approximations. Roughly 35,182 students who took the reading subtest in the 1998-99 school year were classified as failing, despite having an observed score that would have met (or exceeded) the passing criterion had the presence of measurement error been incorporated into the cut score. On the mathematics subtest, 43,077 students who failed met (or exceeded) the adjusted observed criterion score.

Discussion

Because all tests are inherently unreliable to some degree, measurement errors must be accommodated in the development of cut-scores for criterion-referenced tests, particularly when these instruments are used to make high-stakes decisions for student placement. Our analysis focused exclusively on the impact of false negative classification errors. There exists, of course, a second type of misclassification termed a "false positive," or misclassifying *non-passing* students as *passing*. Although both types of misclassification are serious, a survey of Texas educators conducted by Haney (2000) as part of his investigation of the TAAS indicated that respondents viewed the consequences of denying a high school diploma to a qualified student based on a classification error (false negative) as considerably more serious that granting a diploma to an unqualified student (false positive). Indeed, only the consequences to society of granting a license to an unqualified pilot, physician, or teacher, respectively, were viewed as more serious. Additionally, the literature on grade retention is fairly consistent in noting the deleterious effects of these policies (e.g., increased dropout rates), particularly when strong individualized remediation procedures are not in place (McCoy & Reynolds, 1999). Put simply, based on Haney's (2000) survey and the empirical findings on the consequences of retaining students, we feel it seems reasonable to place greater emphasis on the occurrence of false negatives -- at least in the context of education and student promotion decisions.

The estimation process we employed produced results suggesting that about 2% of students who take the state-mandated TAAS exam will be scored as false negatives on one or more of the subtests. The consequences of misclassification will become more evident as the TAAS is increasingly used for promotion and retention decisions in Texas. There is, however, a much larger picture emerging at the national level regarding the (mis)use of standardized assessment tools. To put this in a broader perspective, we extended our analysis to include an estimate of how many students nationally would be misclassified as false negatives if a testing program such as the TAAS were in place. We assumed that testing would include the same grade levels as the Texas accountability model, and assumed also a national testing instrument with the same technical adequacy (reliability) as the TAAS. National data were obtained for the 1998-99 school year disaggregated by grade level. Combining both reading and mathematics error rates results in approximately *1.1 million students* that would potentially be misclassified as false negatives each year across the country. Two percent clearly is no small number in the national context.

The recently installed Bush administration has issued school accountability reform measures that rely almost exclusively on standardized achievement tests. Many questions remain, however, regarding the structure and implementation of the testing program that will serve as a measure of student performance across states. What these tests will look like, the extent to which they yield scores that are psychometrically meaningful, and the importance of the scores in guiding student-level decisions are issues that have not been addressed to date. It is notable that both the American Psychological Association (APA) and, more recently, the American Educational Research Association (AERA) have issued position statements advising against the use of a single assessment for high-stakes decisions at the individual level. It seems probable, however, that the national appetite for school accountability in the form of student achievement scores will overwhelm any concerns over the ethical consequences of high-stakes testing.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando: Harcourt Brace.

Haney, W. (2000). The myth of the Texas Miracle in education. *Education Policy Analysis Archives [On-line serial], 8* (41). Available: *http://epaa.asu.edu/epaa/v8n41/.*

Huynh, H., & Saunders, J. C. (1980). Accuracy of two procedures for estimating reliability of mastery tests. *Journal of Educational Measurement, 17,* 351-358.

Markley, M. . (1999, September 8). HISD rules holding more students back: Expanded standards cut social promotions. *Houston Chronicle*, p. A1.

Messick, S. A. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10,* 1-9.

McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology, 37,* 273-298.

Texas Education Agency (1999, June 22). Briefing book: Legislation affecting public education [On-line]. Available: http://www.tea.state.tx.us/brief/doc2.html.

Wainer, H. (1999). Comments on the ad hoc committee's critique of the Massachusetts Teacher Tests. *Education Policy Analysis Archives [On-line serial], 7* (5). Available: *http://epaa.asu.edu/epaa/v7n5.html.*

Descriptors: Test Validity; Test reliability; Consequences; Impact; Error

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search: title □ Go

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Available online: http://ericae.net/pare/getvn.asp?v=7&n=25. This paper has been viewed 13764 times since 12/11/01.

‣ Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

‣ Find articles in ERIC written by
Mertler, Craig A.

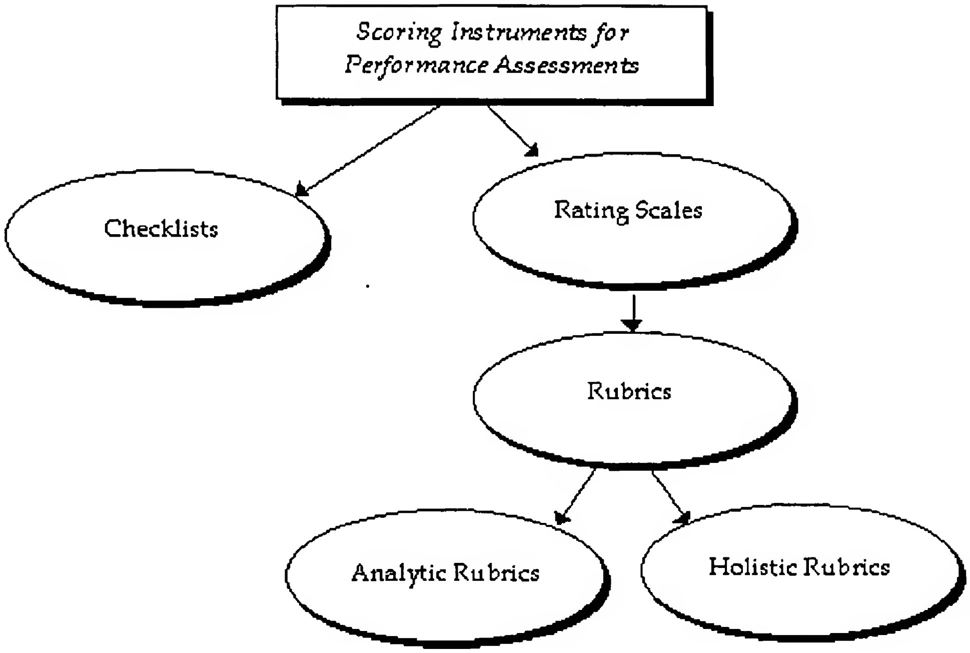
Designing Scoring Rubrics for Your Classroom

Craig A. Mertler
Bowling Green State University

Rubrics are rating scales-as opposed to checklists-that are used with performance assessments. They are formally defined as scoring guides, consisting of specific pre-established performance criteria, used in evaluating student work on performance assessments. Rubrics are typically the specific form of scoring instrument used when evaluating student performances or products resulting from a performance task.

There are two types of rubrics: holistic and analytic (see Figure 1). A **holistic rubric** requires the teacher to score the overall process or product as a whole, without judging the component parts separately (Nitko, 2001). In contrast, with an **analytic rubric**, the teacher scores separate, individual parts of the product or performance first, then sums the individual scores to obtain a total score (Moskal, 2000; Nitko, 2001).

Figure 1:
Types of scoring instruments for performance assessments



Holistic rubrics are customarily utilized when errors in some part of the process can be tolerated provided the overall quality is high (Chase, 1999). Nitko (2001) further states that use of holistic rubrics is probably more appropriate when performance tasks require students to create some sort of response and where there is no definitive correct answer. The focus of a score reported using a holistic rubric is on the overall quality, proficiency, or understanding of the specific content and skills-it involves assessment on a unidimensional level (Mertler, 2001). Use of holistic rubrics can result in a somewhat quicker scoring process than use of analytic rubrics (Nitko, 2001). This is basically due to the fact that the teacher is required to read through or otherwise examine the student product or performance only once, in order to get an "overall" sense of what the student was able to accomplish (Mertler, 2001). Since assessment of the overall performance is the key, holistic rubrics are also typically, though not exclusively, used when the purpose of the performance assessment is summative in nature. At most, only limited feedback is provided to the student as a result of scoring performance tasks in this manner. A template for holistic scoring rubrics is presented in Table 1.

Table 1:
Template for Holistic Rubrics

Score	Description
5	Demonstrates complete understanding of the problem. All requirements of task are included in response.
4	Demonstrates considerable understanding of the problem. All requirements of task are included.
3	Demonstrates partial understanding of the problem. Most requirements of task are included.
2	Demonstrates little understanding of the problem. Many requirements of task are missing.
1	Demonstrates no understanding of the problem.
0	No response/task not attempted.

Analytic rubrics are usually preferred when a fairly focused type of response is required (Nitko, 2001); that is, for performance tasks in which there may be one or two acceptable responses and creativity is not an essential feature of the students' responses. Furthermore, analytic rubrics result initially in several scores, followed by a summed total score-their use represents assessment on a multidimensional level (Mertler, 2001). As previously mentioned, the use of analytic rubrics can cause the scoring process to be substantially slower, mainly because assessing several different skills or characteristics individually requires a teacher to examine the product several times. Both their construction and use can be quite time-consuming. A general rule of thumb is that an individual's work should be examined a separate time for each of the specific performance tasks or scoring criteria (Mertler, 2001). However, the advantage to the use of analytic rubrics is quite substantial. The degree of feedback offered to students-and to teachers-is significant. Students receive specific feedback on their performance with respect to each of the individual scoring criteria-something that does not happen when using holistic rubrics (Nitko, 2001). It is possible to then create a "profile" of specific student strengths and weaknesses (Mertler, 2001). A template for analytic scoring rubrics is presented in Table 2.

Table 2:
Template for analytic rubrics

	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	Score
Criteria #1	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #2	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #3	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #4	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	

Prior to designing a specific rubric, a teacher must decide whether the performance or product will be scored holistically or analytically (Airasian, 2000 & 2001). Regardless of which type of rubric is selected, specific performance criteria and observable indicators must be identified as an initial step to development. The decision regarding the use of a holistic or analytic approach to scoring has several possible implications. The most important of these is that teachers must consider first how they intend to use the results. If an overall, summative score is desired, a holistic scoring approach would be more desirable. In contrast, if formative feedback is the goal, an analytic scoring rubric should be used. It is important to note that one type of rubric is not inherently better than the other-you must find a format that works best for your purposes (Montgomery, 2001). Other implications include the time requirements, the nature of the task itself, and the specific performance criteria being observed.

As you saw demonstrated in the templates (Tables 1 and 2), the various levels of student performance can be defined using either quantitative (i.e., numerical) or qualitative (i.e., descriptive) labels. In some instances, teachers might want to utilize both quantitative and qualitative labels. If a rubric contains four levels of proficiency or understanding on a continuum, quantitative labels would typically range from "1" to "4." When using qualitative labels, teachers have much more flexibility, and can be more creative. A common type of qualitative scale might include the following labels: master, expert, apprentice, and novice. Nearly any type of qualitative scale will suffice, provided it "fits" with the task.

One potentially frustrating aspect of scoring student work with rubrics is the issue of somehow converting them to "grades." It is not a good idea to think of rubrics in terms of percentages (Trice, 2000). For example, if a rubric has six levels (or "points"), a score of 3 should not be equated to 50% (an "F" in most letter grading systems). The process of converting rubric scores to grades or categories is more a process of logic than it is a mathematical one. Trice (2000) suggests that in a rubric scoring system, there are typically more scores at the average and above average categories (i.e., equating to grades of "C" or better) than there are below average categories. For instance, if a rubric consisted of nine score categories, the equivalent grades and categories might look like this:

Table 3:
Sample grades and categories

Rubric Score	*Grade*	*Category*
8	A+	Excellent
7	A	Excellent
6	B+	Good
5	B	Good
4	C+	Fair
3	C	Fair
2	U	Unsatisfactory
1	U	Unsatisfactory
0	U	Unsatisfactory

When converting rubric scores to grades (typical at the secondary level) or descriptive feedback (typical at the elementary level), it is important to remember that there is not necessarily one correct way to accomplish this. The bottom line for classroom teachers is that they must find a system of conversion that works for them and fits comfortably into their individual system of reporting student performance.

Steps in the Design of Scoring Rubrics

A step-by-step process for designing scoring rubrics for classroom use is presented below. Information for these procedures was compiled from various sources (Airasian, 2000 & 2001; Mertler, 2001; Montgomery, 2001; Nitko, 2001; Tombari & Borich, 1999). The steps will be summarized and discussed, followed by presentations of two sample scoring rubrics.

Step 1: *Re-examine the learning objectives to be addressed by the task.* This allows you to match your scoring guide with your objectives and actual instruction.

Step 2: *Identify specific observable attributes that you want to see (as well as those you don't want* *to see) your students demonstrate in their product, process, or performance.* Specify the characteristics, skills, or behaviors that you will be looking for, as well as common mistakes you do not want to see.

Step 3: *Brainstorm characteristics that describe each attribute.* Identify ways to describe above average, average, and below average performance for each observable attribute identified in Step 2.

Step 4a: *For holistic rubrics, write thorough narrative descriptions for excellent work and poor work incorporating each attribute into the description.* Describe the highest and lowest levels of performance combining the descriptors for all attributes.

Step 4b: *For analytic rubrics, write thorough narrative descriptions for excellent work and poor work for each individual attribute.* Describe the highest and lowest levels of performance using the descriptors for each attribute separately.

Step 5a: *For holistic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for the collective attributes.* Write descriptions for all intermediate levels of performance.

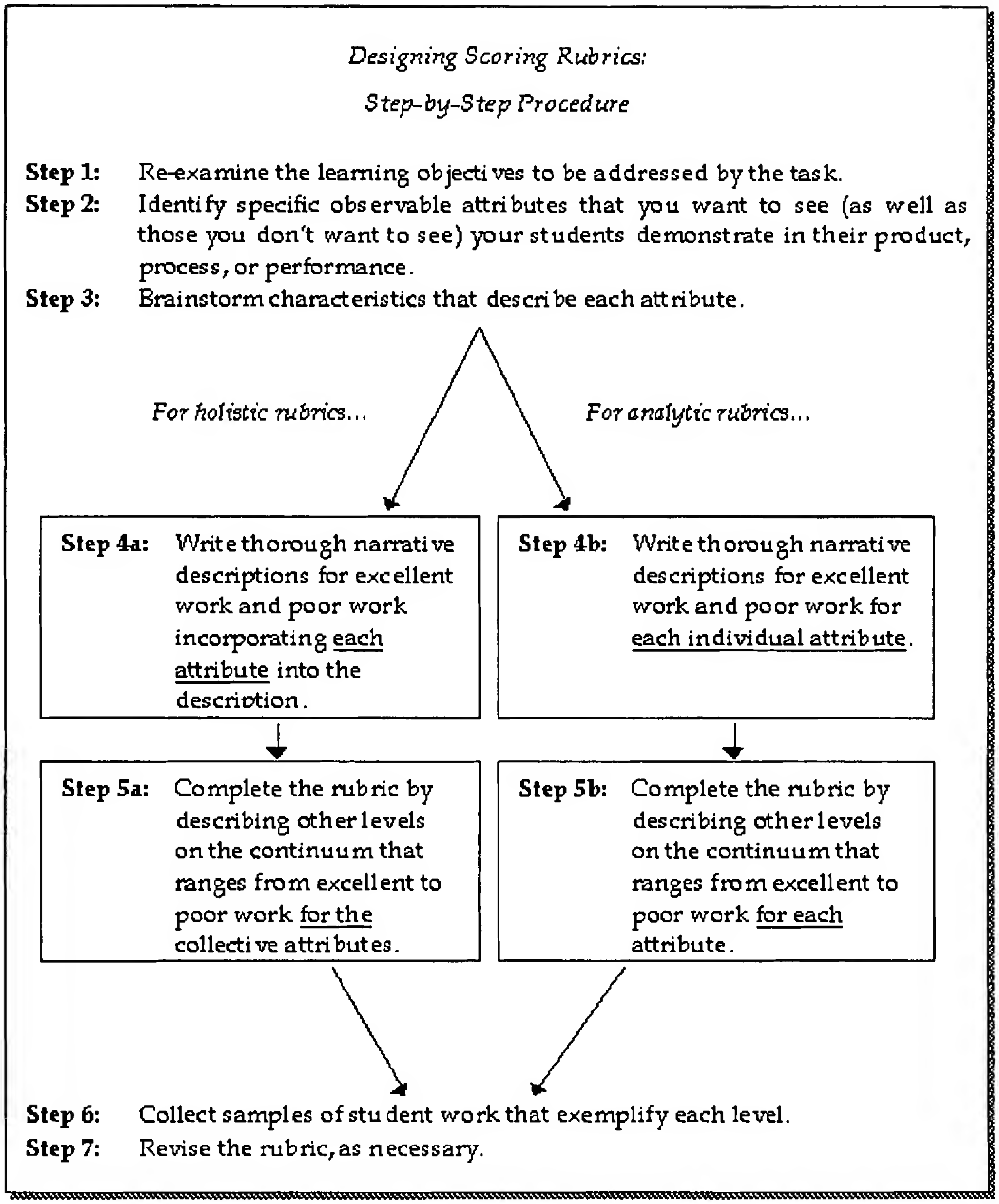
Step 5b: *For analytic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for each attribute.* Write descriptions for all intermediate levels of performance for each attribute separately.

Step 6: *Collect samples of student work that exemplify each level.* These will help you score in the future by serving as benchmarks.

Step 7: *Revise the rubric, as necessary.* Be prepared to reflect on the effectiveness of the rubric and revise it prior to its next implementation.

These steps involved in the design of rubrics have been summarized in Figure 2.

Figure 2:
Designing Scoring Rubrics: Step-by-step procedures



Designing Scoring Rubrics:

Step-by-Step Procedure

Step 1: Re-examine the learning objectives to be addressed by the task.

Step 2: Identify specific observable attributes that you want to see (as well as those you don't want to see) your students demonstrate in their product, process, or performance.

Step 3: Brainstorm characteristics that describe each attribute.

For holistic rubrics...

For analytic rubrics...

Step 4a: Write thorough narrative descriptions for excellent work and poor work incorporating each attribute into the description.

Step 4b: Write thorough narrative descriptions for excellent work and poor work for each individual attribute.

Step 5a: Complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for the collective attributes.

Step 5b: Complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for each attribute.

Step 6: Collect samples of student work that exemplify each level.

Step 7: Revise the rubric, as necessary.

Two Examples

Two sample scoring rubrics corresponding to specific performance assessment tasks are presented next. Brief discussions precede the actual rubrics. For illustrative purposes, a holistic rubric is presented for the first task and an analytic rubric for the second. It should be noted that either a holistic or an analytic rubric could have been designed for either task.

Example 1:
Subject - Mathematics

Grade Level(s) - Upper Elementary

Mr. Harris, a fourth-grade teacher, is planning a unit on the topic of data analysis, focusing primarily on the skills of estimation and interpretation of graphs. Specifically, at the end of this unit, he wants to be able to assess his students' mastery of the following instructional objectives:

- Students will properly interpret a bar graph.
- Students will accurately estimate values from within a bar graph. (step 1)

Since the purpose of his performance task is summative in nature - the results will be incorporated into the students' grades, he decides to develop a holistic rubric. He identifies the following four attributes on which to focus his rubric: estimation, mathematical computation, conclusions, and communication of explanations (steps 2 & 3). Finally, he begins drafting descriptions of the various levels of performance for the observable attributes (steps 4 & 5). The final rubric for his task appears in Table 4.

Table 4:
Math Performance Task – Scoring Rubric
Data Analysis

Name ____________________________ Date __________

Score	Description
4	Makes accurate estimations. Uses appropriate mathematical operations with no mistakes. Draws logical conclusions supported by graph. Sound explanations of thinking.
3	Makes good estimations. Uses appropriate mathematical operations with few mistakes. Draws logical conclusions supported by graph. Good explanations of thinking.
2	Attempts estimations, although many inaccurate. Uses inappropriate mathematical operations, but with no mistakes. Draws conclusions not supported by graph. Offers little explanation.
1	Makes inaccurate estimations. Uses inappropriate mathematical operations. Draws no conclusions related to graph. Offers no explanations of thinking.
0	No response/task not attempted.

Example 2:
Subjects - Social Studies; Probability & Statistics
Grade Level(s) - 9 - 12

Mrs. Wolfe is a high school American government teacher. She is beginning a unit on the electoral process and knows from past years that her students sometimes have difficulty with the concepts of sampling and election polling. She decides to give her students a performance assessment so they can demonstrate their levels of understanding of these concepts. The main idea that she wants to focus on is that samples (surveys) can accurately predict the viewpoints of an entire population. Specifically, she wants to be able to assess her students on the following instructional objectives:

- Students will collect data using appropriate methods.
- Students will accurately analyze and summarize their data.

- Students will effectively communicate their results. (step 1)

Since the purpose of this performance task is formative in nature, she decides to develop an analytic rubric focusing on the following attributes: sampling technique, data collection, statistical analyses, and communication of results (steps 2 & 3). She drafts descriptions of the various levels of performance for the observable attributes (steps 4 & 5). The final rubric for this task appears in Table 5.

Table 5:
Performance Task – Scoring Rubric
Population Sampling

Name ______________________ Date ____________

	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	Score
Sampling Technique	Inappropriate sampling technique used	Appropriate technique used to select sample; major errors in execution	Appropriate technique used to select sample; minor errors in execution	Appropriate technique used to select sample; no errors in procedures	
Survey/ Interview Questions	Inappropriate questions asked to gather needed information	Few pertinent questions asked; data on sample is inadequate	Most pertinent questions asked; data on sample is adequate	All pertinent questions asked; data on sample is complete	
Statistical Analyses	No attempt at summarizing collected data	Attempts analysis of data, but inappropriate procedures	Proper analytical procedures used, but analysis incomplete	All proper analytical procedures used to summarize data	
Communication of Results	Communication of results is incomplete, unorganized, and difficult to follow	Communicates some important information; not organized well enough to support decision	Communicates most of important information; shows support for decision	Communication of results is very thorough; shows insight into how data predicted outcome	

Total Score = ____

Resources for Rubrics on the Web

The following is just a partial list of some Web resources for information about and samples of scoring rubrics.

- "Scoring Rubrics: What, When, & How?" (http://ericae.net/pare/getvn.asp?v=7&n=3). This article appears in Practical Assessment, Research, & Evaluation and is authored by Barbara M. Moskal. The article discusses what rubrics are, and distinguishes between holistic and analytic types. Examples and additional resources are provided.
- "Performance Assessment-Scoring" (http://www.pgcps.pg.k12.md.us/~elc/scoringtasks.html). Staff in the Prince George's County (MD) Public Schools have developed a series of pages that provide descriptions of

the steps involved in the design of performance tasks. This particular page provides several rubric samples.

- "Rubrics from the Staff Room for Ontario Teachers" (http://www.odyssey.on.ca/~elainc.coxon/rubrics.htm) This site is a collection of literally hundreds of teacher-developed rubrics for scoring performance tasks. The rubrics are categorized by subject area and type of task. This is a fantastic resource...check it out!
- "Rubistar Rubric Generator" (http://rubistar.4teachers.org/)
- "Teacher Rubric Maker" (http://www.teach-nology.com/web_tools/rubrics/) These two sites house Web-based rubric generators for teachers. Teachers can customize their own rubrics based on templates on each site. In both cases, rubric templates are organized by subject area and/or type of performance task. These are wonderful resources for teachers!

References

Airasian, P. W. (2000). *Assessment in the classroom: A concise approach* (2nd ed.). Boston: McGraw-Hill.

Airasian, P. W. (2001). *Classroom assessment: Concepts and applications* (4th ed.). Boston: McGraw-Hill.

Chase, C. I. (1999). *Contemporary assessment for educators.* New York: Longman.

Mertler, C. A. (2001). Using performance assessment in your classroom. Unpublished manuscript, Bowling Green State University.

Montgomery, K. (2001). *Authentic assessment: A guide for elementary teachers.* New York: Longman.

Moskal, B. M. (2000). Scoring rubrics: what, when, and how?. *Practical Assessment, Research, & Evaluation,* 7(3). Available online: http://ericae.net/pare/getvn.asp?v=7&n=3

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.

Tombari, M. & Borich, G. (1999). *Authentic assessment in the classroom: Applications and practice.* Upper Saddle River, NJ: Merrill.

Trice, A. D. (2000). *A handbook of classroom assessment.* New York: Longman.

Contact information

Craig A. Mertler
Educational Foundations & Inquiry Program
College of Education & Human Development
Bowling Green State University
Bowling Green, OH 43403

mertler@bgnet.bgsu.edu
Phone: 419-372-9357 Fax: 419-372-8265

Descriptors: *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring

```
ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179
```

Home | Articles | Subscribe | Review | Policies

Volume: 8 7 6 5 4 3 2 1

Practical Assessment, Research & Evaluation
A peer-reviewed electronic journal. ISSN 1531-7714

Search: title □ Go

Rudner, Lawrence & Phill Gagne (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, 7(26). Available online: http://ericae.net/pare/getvn.asp?v=7&n=26. This paper has been viewed 5081 times since 12/12/01.

Find similar papers in
ERICAE Full Text Library
Pract Assess, Res & Eval
ERIC RIE & CIJE 1990-
ERIC On-Demand Docs

Find articles in ERIC written by
Rudner, Lawrence
Phill Gagne

An Overview of Three Approaches to Scoring Written Essays by Computer

Lawrence Rudner and Phill Gagne
University of Maryland, College Park

It is not surprising that extended-response items, typically short essays, are now an integral part of most large-scale assessments. Extended response items provide an opportunity for students to demonstrate a wide range of skills and knowledge, including higher order thinking skills such as synthesis and analysis. Yet assessing students' writing is one of the most expensive and time-consuming activities for assessment programs. Prompts need to be designed, rubrics created, multiple raters need to be trained, and then the extended responses need to be scored, typically by multiple raters. With different people evaluating different essays, interrater reliability becomes an additional concern in the writing assessment process. Even with rigorous training, differences in the background, training, and experience of the raters can lead to subtle but important differences in grading (Blok & de Glopper, 1992, Rudner, 1992).

Computers and artificial intelligence have been proposed as tools to facilitate the evaluation of student essays. In theory, computer scoring can be faster, reduce costs, increase accuracy and eliminate concerns about rater consistency and fatigue. Further, the computer can quickly rescore materials should the scoring rubric be redefined. This articles describes the three most prominent approaches to essay scoring.

Systems

The most prominent writing assessment programs are

- Project Essay Grade (PEG), introduced by Ellis Page in 1966,

- Intelligent Essay Assessor (IEA), first introduced for essay grading in 1997 by Thomas Landauer and Peter Foltz, and
- E-rater, used by Educational Testing Service (ETS) and developed by Jill Burstein.

Descriptions of these approaches can be found at the web sites listed at the end of this article and in Whittington and Hunt (1999) and Wresch (1993). Other software projects are briefly mentioned in Breland and Lytle (1990), Vetterli and Furedy (1997), and Whissel (1994).

Page uses a regression model with surface features of the text (document length, word length, and punctuation) as the independent variables and the essay score as the dependent variable. Landauer's approach is a factor-analytic model of word co-occurrences which emphasizes essay content. Burstein uses a regression model with content features as the independent variables.

PEG - PEG grades essays predominantly on the basis of writing quality (Page, 1966, 1994). The underlying theory is that there are intrinsic qualities to a person's writing style called trins that need to be measured, analogous to true scores in measurement theory. PEG uses approximations of these variables, called proxes, to measure these underlying traits. Specific attributes of writing style, such as average word length, number of semicolons, and word rarity are examples of proxes that can be measured directly by PEG to generate a grade. For a given sample of essays, human raters grade a large number of essays (100 to 400), and determine values for up to 30 proxes. The grades are then entered as the criterion variable in a regression equation with all of the proxes as predictors, and beta weights are computed for each predictor. For the remaining unscored essays, the values of the proxes are found, and those values are then weighted by the betas from the initial analysis to calculate a score for the essay.

Page has over 30 years of research consistently showing exceptionally high correlations. In one study, Page (1994) analyzed samples of 495 and 599 senior essays from the 1998 and 1990 National Assessment of Educational Progress using responses to a question about a recreation opportunity: whether a city government should spend its recreation money fixing up some abandoned railroad tracks or converting an old warehouse to new uses. With 20 variables, PEG reached multiple Rs as high as .87, close to the apparent reliability of the targeted judge groups.

IEA - First patented in 1989, IEA was designed for indexing documents for information retrieval. The underlying idea is to identify which of several calibration documents are most similar to the new document based on the most specific (i.e., least frequent) index terms. For essays, the average grade on the most similar calibration documents is assigned as the computer-generated score (Landauer, Foltz, Laham, 1998).

With IEA, each calibration document is arranged as a column in a matrix. A list of every relevant content term, defined as a word, sentence, or paragraph, that appears in any of the calibration documents is compiled, and these terms become the matrix rows. The value in a given cell of the matrix is an interaction between the presence of the term in the source and the weight assigned to that term. Terms not present in a source are assigned a cell value of 0 for that column. If a term is present, then the term may be weighted in a variety of ways, including a 1 to indicate that it is present, a tally of the number of times the term appears in the source, or some other weight criterion representative of the importance of the term to the document in which it appears or to the content domain overall.

Each essay to be graded is converted into a column vector, with the essay representing a new source with cell values based on the terms (rows) from the original matrix. A similarity score is then calculated for the essay column vector relative to each column of the rubric matrix. The essay's grade is determined by averaging the similarity scores from a predetermined number of sources with which it is most similar. Their system also provides a great deal of diagnostic and evaluative feedback. As with PEG, Foltz, Kintsch and Landauer (1998) also report high correlations between IEA scores and human scored essays.

E-rater - The Educational Testing Service's Electronic Essay Rater (e-rater) is a sophisticated "Hybrid Feature

Technology" that uses syntactic variety, discourse structure (like PEG) and content analysis (like IEA). To measure syntactic variety, e-rater counts the number of complement, subordinate, infinitive, and relative clause and occurrences of modal verbs (would, could) to calculate ratios of these syntactic features per sentence and per essay. For structure analysis, e-rater uses 60 different features, similar to PEG's proxes.

Two indices are created to evaluate the similarity of the target essay's content to the content of calibrated essays. As described by Burstein, et.al (1998), in their *EssayContent* analysis module, the vocabulary of each score category is converted to a single vector whose elements represent the total frequency of each word in the training essays for that holistic score category. The system computes correlations between the vector for a given test essay and the vectors representing the trained categories. The score that is most similar to the test essay is assigned as the evaluation of its content. E-rater's *ArgContent* analysis module is based on the inverse document frequency, like IEA. The word frequency vectors for the score categories are converted to vectors of word weights. Scores on the different components are weighted using regression to predict human grader's scores.

Analysis

Several studies have reported favorably on PEG, IEA, and e-rater. A review of the research on IEA found that its scores typically correlate as well with human raters as the raters do with each other (Chung & O'Neil, 1997). Research on PEG consistently reports relatively high correlations between PEG and human graders relative to correlations between human graders (e.g., Page, Poggio, & Keith, 1997). E-rater was deemed so impressive it is now operational and used to score the General Management Aptitude Test (GMAT). All of the systems return grades that correlate significantly and meaningfully with those of human raters.

Compared to IEA and e-rater, PEG has the advantage of being conceptually simpler and less taxing on computer resources. PEG is also the better choice for evaluating writing style, as IEA returns grades that have literally nothing to do with writing style. IEA and e-rater, however, appear to be the superior choice for grading content, as PEG relies on writing quality to determine grades.

All three of these systems are proprietary and details of the exact process are not generally available. We do not know, for example, what variables are in any model nor their weights. The use of automated essay scoring is also somewhat controversial. A well-written essay about baking a cake could receive a high score if PEG were used to grade essays about causes of the American Civil War. Conceivably, IEA could be tricked into giving a high score to an essay that was a string of relevant words with no sentence structure whatsoever. E-rater appears to overcome some of these criticisms at the expense of being fairly complicated. These criticisms are more problematic for PEG than for IEA and e-rater.

One should not expect perfect accuracy from any automated scoring approaches. The correlation of human ratings on state assessment constructed-response items is typically only .70 - .75. Thus, correlating with human raters as well as human raters correlate with each other is not a very high, nor very meaningful, standard. Because the systems are all based on normative data, the current state of the art does not appear conducive for scoring essays that call for creativity or personal experiences. The greatest chance of success for essay scoring appears to be for long essays that have been calibrated on large numbers of examinees and which have a clear scoring rubric.

Those who are interested in pursuing essay scoring may be interested in the Bayesian Essay Test Scoring sYstem (BETSY), being developed by the author based on the naive Bayes text classification literature (e.g., McCallum and Nigam, 1998). Free software is available for research use.

While recognizing the limitations, perhaps it is time for states and other programs to consider automated scoring services. We don't advocate abolishing human raters. Rather we can envision the use of any of these technologies as a validation tool with each essay scored by one human and by the computer. When the scores differ, the essay would be flagged for a second read. This would be quicker and less expensive than current practice.

We would also like to see retired essay prompts used as instructional tools. The retired essays and grades can be used to calibrate a scoring system. The entire system could then be made available to teachers to help them work with students on writing and high-order skills. The system could also be coupled with a wide range of diagnostic information, such as the information currently available with IEA.

Key web sites

PEG - http://134.68.49.185/pegdemo/ref.asp
IEA - http://www.knowledge-technologies.com/
E-rater - http://www.ets.org/research/erater.html
BETSY - http://ericae.net/betsy/

References and Recommended Reading

Blok, H., & de Glopper, K. (1992). Large-scale writing assessment. In L. Verhoeven (Ed.), J. H. A. L. De Jong (Ed.), *the Construct of Language Proficiency: Applications of Psychological Models to Language Assessment,* pp. 101-111. Amsterdam, Netherlands: John Benjamins Publishing Company.

Breland, H. M., & Lytle, E. G. (1990). Computer-assisted writing skill assessment using WordMAP. ERIC Document Reproduction Service No. ED 317 586.

Burstein, J., K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M.D. Harris (1998). Automated scoring using a hybrid feature identification technique. In the *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, August, 1998. Montreal, Canada. Available on-line: http://www.ets.org/reasearch/aclfinal.pdf

Burstein, J. (1999). Quoted in Ott, C. (May 25, 1999). Essay questions. *Salon.* Available online: http://www.salonmag.com/tech/feature/1999/05/25/computer_grading/

Chung, G. K. W. K., & O'Neil, H. F., Jr. (1997). Methodological Approaches to Online Scoring of Essays. ERIC Document Reproduction Service No. ED 418 101.

Fan, D. P., & Shaffer, C. L. (1990). Use of open-ended essays and computer content analysis to survey college students' knowledge of AIDS. *College Health, 38,* 221-229.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, *25*, (2&3), 285-307.

Jones, B. D. (1999). Computer-rated essays in the English composition classroom. *Journal of Educational Computing Research, 20(*2), 169-187.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211-240.

Landauer, T. K., Foltz, P. W, & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

McCallum, A. and K. Nigam (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-

98 Workshop on "Learning for Text Categorization". Available on-line: http://citeseer.nj.nec.com/mccallum98comparison.html

McCurry, N., & McCurry, A. (1992). Writing assessment for the twenty-first century. *Computer Teacher, 19,* 35-37.

Page, E. B. (1966). Grading essays by computer: Progress report. Notes from the 1966 Invitational Conference on Testing Problems, 87-100.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 127-42.

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). Computer analysis of student essays: Finding trait differences in the student profile. AERA/NCME Symposium on Grading Essays by Computer.

Rudner, L.M. (1992). Reducing errors due to the use of judges. *Practical Assessment, Research & Evaluation*, 3(3). Available online: http://ericae.net/pare/getvn.asp?v=3&n=3.

Vetterli, C. F., & Furedy, J. J. (1997). Correlates of intelligence in computer measured aspects of prose vocabulary: Word length, diversity, and rarity. *Personality and Individual Differences, 22(*6), 933-935.

Whissel, C. (1994). A computer program for the objective analysis of style and emotional connotations of prose: Hemingway, Galsworthy, and Faulkner compared. *Perceptual and Motor Skills, 79*, 815-824.

Whittington, D., & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. *Proceedings of the Third Annual Computer Assisted Assessment Conference,* 207-219. Available online: http://cvu.strath.ac.uk/dave/publications/caa99.html.

Wresch, W. (1993) The Imminence of Grading Essays by Computer - 25 Years Later. *Computers and Composition,* 10(2), 45-58. Available online: http://corax.cwrl.utexas.edu/cac/archiveas/v10/10_2_html/10_2_5_Wresch.html.

Descriptors: Essays; Constructed Response; Scoring; Artifical Intelligence

ADODB.Recordset error '800a0e78'

Operation is not allowed when the object is closed.

/pare/getvn.asp, line 179